كلية علوم الحاسب والمعلومات
قسم هندسة البرمجيات

King Saud University

College of Computer and Information Sciences

Department of Software Engineering

**SWE 486 – Cloud Computing and Big Data**

# Expo 2020 Dubai

## Sentiment Analysis

PHASE 1

---

Project GitHub Repository



| # | NAME | ID | SECTION |
|---|---|---|---|
| 1 | Dalal Bin Humaid | | 68348 |
| 2 | Muneerah Alsunaidi | | 54978 |
| 3 | Reem Aldosari | | 54979 |
| 4 | Sarah alsuqair | | 68348 |
| 5 | Shahad Alshahrani | | 54978 |
| 6 | Safia Assiri | | 54978 |

| | |
|---|---|
| **GROUP #** | 8 |
| **SUPERVISOR** | Hailah Almazrua |



*Submission Date: March 10, 2022*

TABLE OF CONTENTS

TABLE OF FIGURES

# LIST OF TABLES

PART I

# PROJECT DESCRIPTION

In 1851, Expo was invented, the world Expos have provided a platform to showcase the greatest innovations that have shaped the world we live in today. Expos are global events dedicated to finding solutions to fundamental challenges facing humanity by offering a journey inside a chosen theme through engaging and immersive activities. Before Expo 2020, Kazakhstan expo was hosted at Kazakhstan from June 2017 to September 2017, with the expo's theme being "Future Energy". It aimed to create a global debate between countries, nongovernmental organizations, companies, and the general public on the crucial question: "How do we ensure safe and sustainable access to energy for all while reducing CO2 emissions?". Expo 2020 will continue the tradition with the latest technology from around the globe, it is currently hosted by Dubai in the United Arab Emirates from October 1, 2021 to March 31, 2022 which provides a lot of events including: Architecture, Arts and Culture, Business and Entrepreneurship, Food and Beverage, Innovation and Technology, Live Events and Performances, Mobility, and more.

Our main goal in this project is to identify the people's impression of Expo 2020, and if they enjoy and learn from it, or if it’s considered as an ordinary event and find out what they would like to experience by analyzing the data available about it on the Twitter application through Arabic tweets.

We will obtain the tweets from Twitter’s API using Python and its libraries, The results can help Expo management to attract more visitors and provide a better experiment.

## INTRODUCTION

In today’s world, technology has become in every place around us. It has more advancements; the data has grown fast and has a considerable volume, so storing, managing, and processing these data is more complicated. As a result, the term Big Data and Cloud Computing appeared, indicating that big data is a large amount of data with various data types. In contrast, cloud computing stores and access data or software over the internet instead of on a local computer's hard drive or storage. Therefore, cloud computing might be used to supply a variety of services [1].

As we mentioned, Expo 2020 is being hosted in Dubai, UAE, from October 1, 2021 to March 31, 2022. Held once every five years and lasts for six months [2]. It is a festival where everyone can have new experiences, explore, and innovate, and enjoy exchanging ideas and working together. It inspires individuals to showcase the best examples of synergy, innovation, and collaboration from worldwide. Since this is the first time the exhibition is being hosted in the Middle East and West Asia. Engagement has been growing especially from the middle east. In this project, tweets related to Expo 2020 will be extracted from Twitter using Python programming language. Then, we will analyze people's tweets to examine their experiences and opinions about Expo 2020.

## GOAL

During research about Expo 2020 using hashtags such as #اكسبو #اكسبو_دبي, we noticed a significant variation of visitors' opinions.

Our main goal is to measure visitors' satisfaction, find out what the visitor's opinions and interests, what they would like to experience more and what they like or dislike, and find any problems facing them. The results can help Expo's management give them advice or help them attract more visitors and provide a better experiment.

## INITIAL HYPOTHESIS

This section covers the initial hypothesis which is an essential part of the discovery phase. Before we gather and analyze the data, we must formulate an initial hypothesis which we try to approve and disapprove of the following:

IH. Lengthier tweets on a particular activity indicate that it has attracted the attention of visitors in a neutral-positive leaning manner.

H1. Comments about different activities indicate the exhibition's diversity

H2. Most visitors have positive opinions about Expo 2020

## OBJECTIVE

In this project, we aim to analyze Twitter tweets about Expo 2020 and draw conclusions that will help us reach different objectives which are:

1. Measure customer satisfaction
2. Discover common problems that visitors encountered
3. Provide suggestions that increase visitors' satisfaction and attraction
4. Measure visitors' enjoyment and how to increase it
5. Getting people's opinions

# ANALYSIS PLAN

This section covers how we plan to collect our data, what tools and resources we will be using, and how we will coordinate and store the data. At this stage, we have formed our initial hypothesis as discussed in the previous section, which we can build our findings upon and test our progress based on. We have also included extra hypotheses that are merely for curiosity purposes. In this part, we will discuss the technical aspects of gathering the data and what tools/libraries we will be using. The next section will cover the implementation details.

Since our goal is to measure visitors' satisfaction towards Expo 2020. The best platform to do so is on Twitter, the reasons are as follows; the topic is relatively new, using other sources will most likely not reflect actual opinions. Furthermore, Twitter is treated as a personal blog where majority of users reflect their actual feelings/thoughts without any filters or bias. Lastly, people can target public tweets to Expo 2020 which means we have actual insight into what they perceive the exhibition as. Hence, we will use Twitter as our main and only data source.

To gain access to tweets using Twitter API, we need access to the developers' platform. Once that was established, we were equipped with all the essentials to start our data extraction and exploration process.

# DATA EXTRACTION PROCESS

This is the general outline of how we plan to cover this phase and our data collection process.

1. **API connection and configuration**, in this step we will need two libraries. Tweepy to communicate with twitter's API, and configparser to hide API keys and still manage to share the rest of the code. The two libraries will be mentioned in detail in the next part.
2. **Data collection**, once our API setup is complete, we can start gathering the data. Tweepy's search function has a **q** – *query* parameter that enables us to query Twitter for tweets. It is similar to database queries. For instance, *q = 'medical OR automotive'* will result in tweets that either have the term 'medical' or 'automotive' in them. Also, using the following at the end of a query *'-filter-retweets'* will remove any retweets, which allows the data to be somewhat free of duplication. The details of how our query process is performed will be elaborated on in the next section. It is crucial for us to understand how the query can be formed, as it allows us to eliminate unrelated tweets. There are two main approaches we would like to use the query for:

2.1. Query tweets using a defined hashtag/s or keyword/s. This can be querying tweets with *#expo2020* in them or with keyword *expo dubai* and so on. This may include some tweets that are irrelevant for our topics, since many use hashtags that are unrelated to gain extra exposure. But the majority will hopefully be related.

2.2. Query tweets with a specific mention. In our case, Expo 2020 has an account *@expo2020dubai*, we can fetch all the tweets that mentioned it. As previously

mentioned, this can help us gain insight as to what exactly people are perceiving the exhibition as.

3. **Data extraction**, in this part, we will explore the data to make sure it is accurate and relevant and most importantly is stored correctly. We can then plot the data and start exploring certain aspects to ensure everything aligns with our vision and hypotheses. This can consist of counting the number of tweets, plotting it against time. The goal is to gain extra knowledge about the data and familiarize ourselves with it. We can use libraries such as matplotlib.
4. **Data storing**, once we are satisfied with the amount of data we have collected, we can store it in a .csv file to use later.

## DEVELOPMENT PLATFORMS

1. **Python**
   An interpreted, high level programming language that gained huge attraction in the fields of data science and machine learning [3]. It is widely versatile and contains libraries that support image processing and augmentation, text manipulation, statistical tools, and visualization. We will use many Python built-in and community libraries to complete the project.
2. **Anaconda**
   A distribution environment for languages like Python [4]. it is essentially a collection of related files, packages and tools that ease the process of data analysis. It is a one stop shop once you install anaconda you will have access to many libraries and packages that otherwise you would have to install manually. A huge drawback, at least to us is the fact that it takes huge memory storage[1]. For this reason, some of the team members preferred a different route to setup their environment.
3. **Jupyter Notebook**
   An interactive Python notebook. It allows us to write in markdown and add code blocks in the same document. Another feature is the ability to run different code blocks independently. Not only is it formatted in a readable way it can also be exported in many extensions from HTML to pdf and even markdown. The notebook can run locally and or be hosted on a server. when running the command *jupytr notebook* it will launch the folder directory on a local host website.
4. **VS Code**
   Some team members opted to not use Anaconda and Jupyter, instead they have installed all the needed libraries and worked inside of Visual Studio Code. It is preferable since they will be familiar with the shortcuts and the general interface. Also, there is much more control with the installation and debugging process.

It is important the team gets acquainted with their setup and environment. Which is why this section is being elaborated on to such an extent. It is also crucial to learn or brush-up on Python's syntax. Since we will work with it throughout the entirety of the project.

[1] There is another distribution called miniconda which aims to resolve this problem

## TOOLS

These are the tools and libraries we will be using during this phase and the rest of the project. We will briefly describe them and what they will be used for.

1. **Tweepy**
   An open-source Python package that provides a very convenient way to access the Twitter API with Python, it includes a set of classes and methods that represent Twitter's models and API endpoints, and it transparently handles various implementation details [5].
2. **Configparser**
   A Python class which implements a basic configuration language for Python programs, it can provide a structure similar to Microsoft Windows INI files and allows to write Python programs which can be customized by end users easily [6]. We used it to mask and hide our API keys, the *.ini* extension was added to *.getignore*. This allowed us to share the code via GitHub and collaborate without affecting our API details.
3. **Pandas**
   A Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labelled" data both easy and intuitive [7].
4. **Matplotlib**
   A cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open-source alternative to MATLAB. Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications [8]. We will dive deeper with matplotlib in the upcoming phases. However, we did a simple visualization to get ourselves acquainted during this phase.
5. **Random**
   An in-built module of Python which is used to generate random numbers. These are pseudo-random numbers means these are not truly random [9]. This module can be used to perform random actions such as generating random numbers, print random a value for a list or string, etc. In our case we used it to extend .csv files since we will be running the method multiple times and from different PC's its more convenient to append a random number than keeping track of the file numberings.

# DATA COLLECTION

Twitter API will be used as a data source without any additional dataset. The following steps describe the process of data extraction and collection. It's important to note that each step of the code is described in its own markdown. This way it's easier for others to follow and run the code locally. Here, will go through each step in greater detail.

## THE PROCESS

1. **Import Libraries** and tools mentioned in the previous section

```
import tweepy
import pandas as pd
import numpy as np
import configparser
import matplotlib.pyplot as plt
import random
✓ 3.7s
```

*Figure 1 Libraries and tools import*

2. **Connect to Twitter's API**

2.1. Generate Consumer API and Access keys and store them in 'config.ini' file.

```
[twitter]
CONSUMER_KEY = 'YOUR CONSUMER KEY'
CONSUMER_SECRET = 'YOUR CONSUMER SECRET'
ACCESS_TOKEN = 'YOUR ACCESS TOKEN'
ACCESS_TOKEN_SECRET = 'YOUR ACCESS TOKEN SECRET'
```

*Figure 2 'config.ini' file format*

```
# read the file from 'config.ini'
config = configparser.ConfigParser()
config.read('config.ini')

# API Variables
CONSUMER_KEY = config['twitter']['CONSUMER_KEY']
CONSUMER_SECRET = config['twitter']['CONSUMER_SECRET']
ACCESS_TOKEN = config['twitter']['ACCESS_TOKEN']
ACCESS_TOKEN_SECRET = config['twitter']['ACCESS_TOKEN_SECRET']
```

*Figure 3 Read the API keys*

2.2. Authenticate access using tweepy, after obtaining keys from ‘config.ini’ file.

```python
# authenticate using tweepy
def twitter_setup():
    auth = tweepy.OAuth1UserHandler(CONSUMER_KEY, CONSUMER_SECRET)  # project access
    auth.set_access_token(ACCESS_TOKEN, ACCESS_TOKEN_SECRET)  # user access

    api = tweepy.API(auth = auth)
    return api

extractor = twitter_setup()
```

*Figure 4 Tweepy authentication setup*

3. **Extract Tweets** and store them into a .csv file.
We went through multiple iterations for writing an extraction function that does the job efficiently and collects as much data as possible. The first iteration we followed an ad-hoc approach by simply writing the query and trying any possible combination of words/hashtags. We also made a mistake by setting the .items() to 1000. We assumed that the higher the value the more data we would receive. Unfortunately, twitter has rate limits which in our case was 450 tweets/15-minute window, which we had encountered a lot.

Upon creating multiple iterations of the previous approach, we decided that we should first create a list of search terms labeled as *search_terms* that we would iterate through each time and query the API based on. The following is what we decided the search terms should cover:

1. Any mention of Expo 2020’s account *@expo2020dubai*
2. Tweets with the hashtags *#expo2020* or #اكسبو
3. Tweets that have both Expo and Dubai in them (in Arabic)

We had a longer list, but we noticed many of them overlap which produced the exact same tweets. We then added a *-filter:retweets* to further remove any redundancy. Also, we filtered the search to only fetch Arabic tweets.

Afterwards, we simply fetched the tweets given the term and store them in a data frame that is then merged with the previously generated one and the iteration continues. We finally store them in a .csv file and returned the final data frame.

This approach fetches around 1900 tweets each time it is called. The biggest drawback or limitation we faced is that the *search_tweets* API only fetched results within the last 7 days. We did try to gain access to the premium search API in order to search the full archive but we did not have any luck. With that being said, we luckily started fetching the data early. Therefore, we have some time variation within the collected tweets. **See figure 5** for the full function.

```python
def extract_tweets():
    tweets = []  # main data frame
    data = [] # temporary data frame
    columns_header = ['ID', 'Tweet', 'Timestamp', 'Likes', 'Retweets', 'Length']
    search_terms = ['@expo2020dubai -filter:retweets',
                    '#expo2020 -filter:retweets',
                    '#اكسبو -filter:retweets',
                    'اكسبو دبي -filter:retweets']  # search terms

    # fetch the tweets once prior to the iteration to append things correctly
    collected_tweets = tweepy.Cursor(extractor.search_tweets, q='expo dubai -filter:retweets', lang='ar', tweet_mode='extended').items(600)

    for tweet in collected_tweets:
        data.append([tweet.id, tweet.full_text, tweet.created_at,tweet.favorite_count, tweet.retweet_count, len(tweet.full_text)])

    tweets = pd.DataFrame(data=data, columns=columns_header) # store in original data frame

    for term in search_terms:
        data = []
        collected_tweets = tweepy.Cursor(extractor.search_tweets, q=term, lang='ar', tweet_mode='extended').items(600)

        for tweet in collected_tweets:
            data.append([tweet.id, tweet.full_text, tweet.created_at, tweet.favorite_count, tweet.retweet_count, len(tweet.full_text)])

        df = pd.DataFrame(data=data, columns=columns_header)
        frames = [tweets, df]
        tweets = pd.concat(frames) # append the data frame to the previous one

    # since we are appending data frames the index value changes each time
    # here the goal is to create a new index that is incremented by one
    tweets.insert(0, 'index', range(0, len(tweets)))
    tweets = tweets.set_index('index')

    # random number to ensure files don't get overwritten
    tweets.to_csv(f'tweets{random.randint(127,1862)}.csv')

    return tweets
```

*Figure 5 Tweet extraction function*

```
tweets = extract_tweets()
✓ 1m 29.1s
```

*Figure 6 Time it took to extract the tweets*

4. **Preliminary Data Exploration,** which consists of retrieving information about the data and displaying it.

```
display(tweets.head())
display(tweets.tail())
print('total of collected tweets is ', len(tweets))
```

✓ 0.1s

| index | ID | Tweet | Timestamp | Likes | Retweets | Length |
|---|---|---|---|---|---|---|
| 0 | 1500939349889884166 | أثناء زيارتها لـ #إكسبو2020 #دبي.. وزيرة تمساو... | 2022-03-07 21:00:01+00:00 | 0 | 0 | 201 |
| 1 | 1500900614473138181 | #RobertoCavalli #Uomo \n@roberto_cavalli \n\n*... | 2022-03-07 18:26:06+00:00 | 0 | 0 | 199 |
| 2 | 1500770169651707904 | ... هذو لي راك تهدر عليهم هم نفسهم الذين اردتم ان | 2022-03-07 09:47:46+00:00 | 1 | 0 | 203 |
| 3 | 1500745009611657224 | مع عبارة "لن أنساك. نقدم للزوار أ Shion تتوافق... | 2022-03-07 08:07:47+00:00 | 1 | 0 | 285 |
| 4 | 1500730516831715328 | جناح #أنجولا يحتفي مع جمهور #إكسبو2020 #دبي با... | 2022-03-07 07:10:12+00:00 | 1 | 1 | 224 |

| index | ID | Tweet | Timestamp | Likes | Retweets | Length |
|---|---|---|---|---|---|---|
| 1886 | 1500048479569207297 | @zaydoun النجاح اللي حققه جناح الكويت في اكسبو... | 2022-03-05 10:00:01+00:00 | 6 | 0 | 277 |
| 1887 | 1500047308750270464 | #إكسبو2020 #دبي يستضيف فعالية "استثمر في #فلسط... | 2022-03-05 09:55:22+00:00 | 1 | 2 | 117 |
| 1888 | 1500047187828477957 | #القوات_المسلحة_الإماراتية تقدم عرضًا عسكريًا ... | 2022-03-05 09:54:53+00:00 | 0 | 0 | 146 |
| 1889 | 1500046903177891842 | 💔ياليتني كنت معاكم امس في حفل ميحد حمد \n#ميحد... | 2022-03-05 09:53:45+00:00 | 2 | 1 | 90 |
| 1890 | 1500046581567205378 | أنا من يومين عفست الدنيا وخربت جدول ويكند خوات... | 2022-03-05 09:52:29+00:00 | 1 | 0 | 129 |

total of collected tweets is 1891

*Figure 7 Preliminary data exploration and printing the total tweets*

```
tweets.info()
```

[11] ✓ 0.1s

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1891 entries, 0 to 1890
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   ID         1891 non-null   int64
 1   Tweet      1891 non-null   object
 2   Timestamp  1891 non-null   datetime64[ns, UTC]
 3   Likes      1891 non-null   int64
 4   Retweets   1891 non-null   int64
 5   Length     1891 non-null   int64
dtypes: datetime64[ns, UTC](1), int64(4), object(1)
memory usage: 103.4+ KB
```

*Figure 8 Preliminary data exploration, details of the data frame*

## 5. **Visualize Data with Matplotlib**

```
plt.plot(tweets['Timestamp'], tweets['Likes'])
plt.gcf().autofmt_xdate()

plt.show()
```

[12] ✓ 0.3s

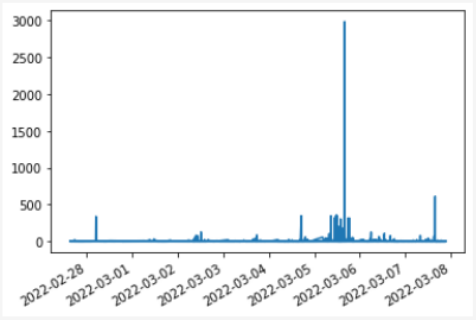


*Figure 9 Simple data visualization I*

```
plt.scatter(tweets['Length'], tweets['Likes'], color='pink')
plt.show()
```

2] ✓ 0.2s

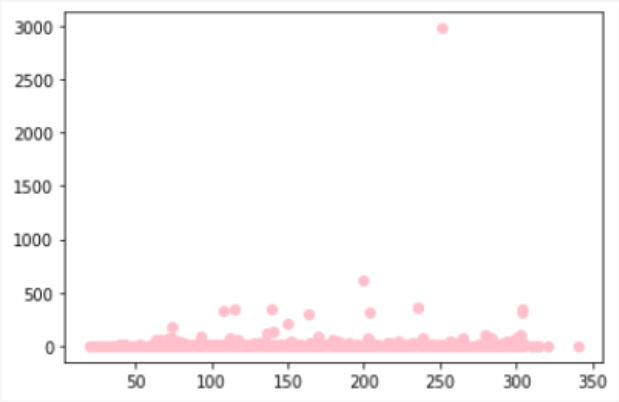


*Figure 10 Simple data visualization II*

## GATHERED DATA

A total of 7767 tweets were collected. We notice that most of the tweets of Expo 2020 are regarding people's experiences, announcements of new events, and people's views and opinions on activities. At first sight, based on the number of likes, emoticons, and the content of tweets, we could perhaps conclude that most people are enthusiastic about Expo 2020. A sample of the data is shown in **figure 11** below.

| index | ID | Tweet | Timestamp | Likes | Retweets | Length |
|---|---|---|---|---|---|---|
| 1875 | 1.50003E+18 | امس تلتهت لدرجه شرطي شافني قاعده ف شارع ويسال اذا فيه شي واخر شي بابلي ماي 😳😳😳 لاتنه نسيت سويج سياره عذ | 3/5/2022 11:49 | 0 | 0 | 145 |
| 1874 | 1.50003E+18 | •••تنظّم مؤسسة محمد بن راشد آل مكتوم للمعرفة بالتعاون مع برنامج الأمم المتحدة الإنمائي، الدورة السابعة من قمة المعرفة | 3/5/2022 11:52 | 20 | 10 | 297 |
| 1873 | 1.50003E+18 | من قلب #اكسبوالله يا دار زايد كيف محلاها #إكسبو_2020_دبي#إكسبو2020 https://t.co/QZrL8eJvGr#ɔ2020Dubai | 3/5/2022 11:53 | 4 | 3 | 123 |
| 1872 | 1.50003E+18 | @expo2020dubai@https://t.co/DYntPD1hFjخاص- فايز السعيد يتألق بليلة ساحرة ضمن أمسيات خالدة في إكسبو دبي | 3/5/2022 11:57 | 2 | 1 | 194 |
| 1871 | 1.50004E+18 | @9xm_9 معي عن اكسبو دبي اذا تريديه | 3/5/2022 12:09 | 5 | 0 | 34 |
| 1870 | 1.50004E+18 | دار زايد #ميحد_حمد اكسبو دبي 2020 @YouTube via https://t.co/zbxJAlttL | 3/5/2022 12:14 | 3 | 1 | 70 |
| 1869 | 1.50004E+18 | قدّمت القوات المسلحة الإماراتية عرضًا عسكريًا في جنوب إكسبو 2020 دبي. إليكم بعض اللقطات وأبرز الأحداث من عرض "حد | 3/5/2022 12:23 | 0 | 0 | 171 |
| 1868 | 1.50004E+18 | مقارنة بسيطة بين تنظيم معرض الكتاب و اكسبو دبي تكشف لك مدى الفرق الكبير بين البلدين، وتبين الفرق بين المسؤولين اذا ك | 3/5/2022 12:31 | 5 | 1 | 207 |
| 1867 | 1.50004E+18 | #المغرب MA، سبو 2020 - دبي" عن رغبته في المساهمة في الجهود المبذولة لمواجهة التحديات العالمية من خلال رؤيته للتنمية | 3/5/2022 12:35 | 0 | 0 | 172 |
| 1866 | 1.50005E+18 | اليوم.. انطلاق "#حصن_الاتحاد8" في #إكسبو_2020_دبي والذي يقام تحت رعاية #محمد_بن_راشد، لإظهار القدرات المتطور | 3/5/2022 12:47 | 62 | 40 | 238 |
| 1865 | 1.50005E+18 | البوثات العالمية في معرض اكسبو دبي كل الدول حاطه المشاريع التنموية لها الا الكويت تتغنى بماضيها #الكويت | 3/5/2022 12:49 | 0 | 1 | 104 |
| 1864 | 1.50005E+18 | ɴttps://t.co/YzlptL2zHX #أخبار_الإمارات #دبي https://t.co/cpq9oBqJ27 سيارة الرجل الوطواط تحلق في فضاء إكسبو | 3/5/2022 12:52 | 0 | 0 | 108 |
| 1863 | 1.50005E+18 | مكار أنا من يومين عفست الدنيا وخربت جدول ويكند خواتي عشان يودوني إكسبو اليوم، ويوم رتبنا كل شي توصلني دعوة غداء في | 3/5/2022 12:52 | 1 | 0 | 129 |
| 1862 | 1.50005E+18 | https://t.co/KUb0HGRYj7 2020_دبي_اكسبو #ميحد_حمد 💔 ياليتني كنت معاكم امس في حفل ميحد حمد | 3/5/2022 12:53 | 2 | 1 | 90 |
| 1861 | 1.50005E+18 | #إكسبو_2020_دبي# أخبار_سعادة#expo2020dubai .القوات_المسلحة_الإماراتية تقدم عرضًا عسكريًا متميزاً في جنوب | 3/5/2022 12:54 | 0 | 0 | 146 |
| 1860 | 1.50005E+18 | https:/ @expo2020dubai #Expo2020#عينك_على_إكسبو# "إكسبو2020 #دبي يستضيف فعالية "استثمر في #فلسطين# | 3/5/2022 12:55 | 1 | 2 | 117 |
| 1859 | 1.50005E+18 | @zaydoun ظ الناس الى رؤساء الدول.جناحكم النجاح اللي حققه جناح الكويت في اكسبو دبي ما يخضع لمبدأ جناحنا احسن من | 3/5/2022 13:00 | 6 | 0 | 277 |
| 1858 | 1.50005E+18 | واصل إكسبو 2020 دبي تعزيز حضوره كأهم وأضخم الأحداث العالمية استقطاباً للزوار والمسؤولين العالميين وصناع القرار تحت | 3/5/2022 13:00 | 1 | 0 | 235 |
| 1857 | 1.50005E+18 | أرض الإمارات الطيّبة، منبت الأبطال وواحة السلام والتسامح، ستكون مسرحاً لعروض حصن الاتحاد 8 في 5 و6 مارس 2022، جـ | 3/5/2022 13:03 | 1 | 3 | 243 |
| 1856 | 1.50005E+18 | #Expo2020Dubai أتوقع اليوم رقم قياسي جديد لزوار أكسبو دبي .. زحمة زحمة | 3/5/2022 13:05 | 0 | 1 | 75 |
| 1855 | 1.50005E+18 | وكيل وزارة الاعلام #منيرة_الهويدي اولاً: مسّاح الله بالخير ثانياً: سؤال معالي الوكيلة الفنانين موظفين بوزارة الاعلام ؟ وشنو دوره | 3/5/2022 13:08 | 3 | 0 | 297 |
| 1854 | 1.50005E+18 | @zaydoun شتان بين الاثنين علما هناك ايضا دول اقل منا وبسيطه ولكن جميله وتشرحانا زرت جناح اكسبو ميلانو واكسبو دبي | 3/5/2022 13:09 | 1 | 0 | 210 |
| 1853 | 1.50005E+18 | محمد بن راشد، ومكتوم بن محمد، خلال زيارتهما السابقة للجناح الجزائري DZ بمعرض إكسبو 2020 دبي. يُذكر أن جناحنا يُبرز | 3/5/2022 13:11 | 0 | 0 | 276 |
| 1852 | 1.50005E+18 | ttps://t.co/Me3Zdkh7aW دبي 2020 ▶ الالتقاء بعدد من الرؤساء التنفيذيين ضمن مشاركة سلطنة عمان في معرض إكسبو | 3/5/2022 13:12 | 8 | 4 | 146 |
| 1851 | 1.50005E+18 | @smralmazrooei1 ول يدخل من ضمن تركيبات الخلطه 🤣. نفس اللي يرفع سعر الشاي في البحر ويقولك بسبب اكسبو دبي | 3/5/2022 13:12 | 0 | 0 | 108 |
| 1850 | 1.50005E+18 | معالي وكيل وزارة الاعلام ... #منيرة_الهويدي انتظري التغريدة القادمة .. عاد شوفي متى انزلها يمكن يمكن فيها "تنف شوارب" ص | 3/5/2022 13:13 | 4 | 0 | 203 |
| 1849 | 1.50005E+18 | ما شاء الله حصل مصرف الإمارات المركزي على جائزة الجهة الاتحادية الرائدة في فئة أفضل جهة تحسنًا في الأداء بالدورة السادسة | 3/5/2022 13:16 | 1 | 0 | 248 |
| 1848 | 1.50005E+18 | ليش ماخبرتوني باركنات اكسبو فيها مليون اسم ولون وانا بس حافظة الحرف امس تميت ساعتين احوط احس شكرا شرطة دبي لو | 3/5/2022 13:19 | 5 | 0 | 159 |
| 1847 | 1.50005E+18 | تحت رعاية #محمد_بن_راشد.. انطلاق العرض العسكري "#حصن_الاتحاد8" اليوم في تمام الساعة 4:30 مساءً وعلى مدى يومٍ | 3/5/2022 13:20 | 13 | 35 | 237 |
| 1846 | 1.50006E+18 | @HadeeLBuQrais !! مارحت اكسبو دبي بس زميلة لي رايحة ومصورة بصراحة شي خيال ومبهر | 3/5/2022 13:27 | 0 | 0 | 79 |
| 1845 | 1.50006E+18 | كة في #إكسبو_2020 في دبي، حيث كان واحدًا من أجمل الأجنحة للدول المشاركة وعددها أكثر من 190 دولةاقتباسات_وطنية# | 3/5/2022 13:31 | 0 | 3 | 228 |
| 1844 | 1.50006E+18 | ٩DdFRhn3 كشف قائمة أسماء 13 عنصرا للموساد الإسرائيلي ينشطون في إكسبو دبي لأغراض استخبارية #الامارات #إكسبو_دبي | 3/5/2022 13:34 | 2 | 2 | 127 |
| 1843 | 1.50006E+18 | https://t.co/D5dSXw0oAG #دبي #كل_الحب #اكسبو_٢٠٢٠ من دبي الجميلة 💚 | 3/5/2022 13:47 | 324 | 11 | 108 |
| 1842 | 1.50006E+18 | إكسبو# "البحث قضية دعم المساواة بين الجنسين في الإبصار".. جناح المرأة في #إكسبو2020 #دبي يستضيف فعالية "هي ترى" | 3/5/2022 13:48 | 2 | 2 | 170 |
| 1841 | 1.50006E+18 | لا حدود للمفاجآت في إكسبو 2020 دبي حيث استحضر الحدث الدولي الكبير سيارة الرجل الوطواط الظاهرة في فيلم "ذا باتمان" | 3/5/2022 13:52 | 1 | 0 | 300 |

*Figure 11 Resulted data sample*

# REFERENCES


[1] [Online]. Available: https://www.computer.org/publications/tech-news/trends/big-data-and-cloud-computing.

[2] [Online]. Available: https://www.expo2020dubai.com/.

[3] "Welcom to Python.org," python.og, 3 March 2022. [Online]. Available: https://www.python.org/. [Accessed 7 March 2022].

[4] "Anaconda," Anaconda Inc., 2018. [Online]. Available: https://www.anaconda.com/. [Accessed 7 March 2022].

[5] "Tweepy Documentation," Tweepy.org, [Online]. Available: https://docs.tweepy.org/en/stable/. [Accessed 7 March 2022].

[6] "configparser — Configuration file parser," Python.org, March 2022. [Online]. Available: https://docs.python.org/3/library/configparser.html. [Accessed 7 March 2022].

[7] "pandas - Python Data Analysis Library," Pydata.org, 2022. [Online]. Available: https://pandas.pydata.org/. [Accessed 7 March 2022].

[8] Matplotlib.org, "Matplotlib: Visualization with Python," 2022. [Online]. Available: https://matplotlib.org/. [Accessed 7 March 2022].

[9] "random — Generate pseudo-random numbers," Pyhton.org, [Online]. Available: https://docs.python.org/3/library/random.html. [Accessed 7 March 2022].

# Appendices

## Files Mapping

*Table 1 Files and their descriptions*

| FILE NAME | DESCRIPTION |
|---|---|
| tweets-collection.ipynb | The data collection Jupyter notebook. You will find a step-by-step process of how to run and execute the data collection procedure. |
| tweets.csv | The full dataset of collected tweets. You will find 7767 entries. |
| tweets.xlsx | The previous dataset but converted to an excel sheet. if you'd like to preview the dataset you should open this file. |